

A Novel Approach for Identification of Hadoop Cloud Temporal Patterns Using Map Reduce

P.Srinivasa Rao

Sr.Asst.Professor, Dept.of CSE, MVGRCE, Vizianagaram, India
E-mail: psr.sri@gmail.com

Dr. K.Thammi Reddy

Professor, Dept.of CSE, GITAM University, Visakhapatnam, India
E-mail: thammireddy@yahoo.com

Dr. MHM.Krishna Prasad

Associate Professor Dept.of CSE, JNTUK, Kakinda, India
E-mail: krishnaprasad.mhm@gmail.com

Abstract– Due to the latest developments in the area of science and Technology resulted in the developments of efficient data transfer, capability of handling huge data and the retrieval of data efficiently. Since the data that is stored is increasing voluminously, methods to retrieve relative information and security related concerns are to be addressed efficiently to secure this bulk data. Also with emerging concepts of big data, these security issues are a challenging task. This paper addresses the issue of secure data transfer using the concepts of data mining in cloud environment using hadoop mapreduce. Based on the experimentation done results are analyzed and represented with respect to time and space complexity when compared hadoop with non hadoop approach.

Index Terms– Big Data, Hadoop, Mapreduce, Cloud Computing, Temporal Patterns

I. Introduction

The present day technological developments witnessed the storage of huge data and methodologies targeted towards efficient retrievals. Since this data is available are surmounting, security breaches and upholding the privacy is a major concern. These security issues are much more challenging while considering the data transfers in cloud environment or parallel processing architectures [1]. In order to handle this data efficiently concepts of mapreduce [2] is focused in the literature. This is due to its capabilities of faulttolerance and scalability together with simplicity. Another main advantage of highlighting the mapreduce concept is it facilitates the parallel processing environments which help indirectly towards huge data storage [3]. The concept of mapreduce can be easily implemented using hadoop environment [4]. Many

methodologies have been discussed in literature [5, 6, 7, 8] to address the issues of security in client server environment. However among the limited algorithms used for security in distributed environments Symmetric Encryption is mainly projected due to its robustness and capability [12] of usage in both 64 bit and 128 bit key format.

In the present day scenario due to the increase in the cost of software, maintenance of software, storage of software, forced the users or developers to adopt cloud computing environment. In this environment the software or data is stored primarily in the form of clusters. These clusters will be transmitted over a cloud based on the users request types which can be SAS, PAS and IAS [8]. Among the different services provided by the cloud environment, the frequently used services include providing instances on demand and providing computational capabilities on demand. The map reduce concept addressed in this paper supports the distributed computing for large data sets on clusters of computers for providing computing capability on demand. To facilitate this service hadoop is mainly used due to its capability of handling HDFS files by which data related to different machines along the globe can be stored. Mapreduce is a functionality of hadoop which helps in data preprocessing. This preprocessed data can be helpful for the efficient analysis of bigdata. Data mining is the exploration of data with the goal of discovering hidden structure.

In many real-world applications, it is important to study the change of temporal features of a non-stationary time series, and identify the ones that are representing the significance of time instances. For example, it is critical in data leakage applications from where the data has been leaked or it is difficult to identify IP of an unauthorized user who logged at any time or irregular interval of time in a cloud environment generally such time series are considered

non-stationary. Traditional time series analysis employs statistical methods to model and explain the data and predict future values of the time series. It is not easy, however, to identify the critical temporal patterns of the time series using these traditional methods. Using a set of observations, in this paper, we present a new method for time series data mining. By incorporating symmetric key encryption with the use of hadoop, temporal patterns (user's log history at regular or irregular time interval) can be effectively revealed in non-stationary (cloud) environment. In order to handle the huge data and transmit the data across the globe efficient data transfer methodologies are to be adopted by applying symmetric encryption.

1.1 Reasons for Symmetric Encryption

However among the limited algorithms used for security in distributed environments Symmetric Encryption is mainly projected due to its robustness and capability [12] of usage in both 64 bit and 128 bit key format.

1.2 Non-Hadoop Approach

The current processing of data goes through ordinary sequential ways to accomplish this job. The program loads data, processing each data block alone before writing the newly processed data block on a storage device. Many ordinary tools available, for example. Besides, many ordinary C and Java programs can be downloaded from the Internet or easily developed to perform such tasks. Most of these tools run on a single computer with a Windows operating system. Although batch processing can be found in these single-processor programs, there will be problems with the processing due to limited capabilities. Therefore, we are in need of a new parallel approach to work effectively on massed data which should be transmitted securely in cloud environment.

1.3 Contribution and Plan of the Paper

In this paper, we are going to address security related issues of the data that has been exchanged between Hadoop machines in Distributed Environment. The rest of this paper is organized as follows. In section2, the related work is discussed. In section3, proposed system model and architecture was presented. In section4 the Methodology is presented. In section5 experimentation results are presented. Finally Conclusions and Future work are made in section6.

II. Related Work

P.Srinivasa Rao et al[18] proposed a method to protect internet usage from unauthorized users by using hadoop mapreduce where a namenode log file approach

is proposed in which identification of user's temporal patterns approach experimented. ElisaBertino et al[10] proposed a method of Digital Identity Management for a cloud using Multifactor Authentication Technique .S.Fiseher-Hubnar et al[11, 15, 16] proposed a privacy and Identity Management for Europe where it provides Privacy Preservation Authentication using Erroneous Credentials.Basker Prasad Rimal et al[8]. proposed a method in understanding of taxonomy and survey of cloud computing systems Kumar Gunjan et al[13] gave an overview idea of identity management in cloud computing Mark D.Ryan et al[14, 17] explained cloud computing security: the scientific challenge and survey of solutions . Taking into account all the above issues, in this paper I am going to address Protection of huge data that has been transmitted in cloud through Hadoop Distributed System by using DES Algorithm. To minimize the delay due to decryption process at the receiving end, an alternative approach can be adopted for authorized users by sending raw data. In this paper i propose a novel methodology where in the security will be provided in two stages in Hadoop Cloud environment.

III. System Model

In the Figure1 illustrated above, at any slave(datanode) user may send the data to any other datanode in the cloud. The process of sending the data securely to other destination node is clearly visualized in the above figure.

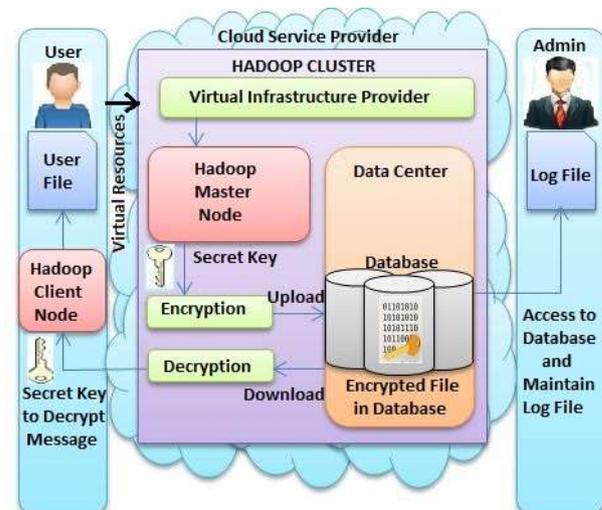


Fig. 1: Architecture of Hadoop Cluster in Cloud

3.1 Internal operation

At any datanode if the user is getting authorization to enter into cloud, he can be allowed to send or receive data that can be processed is shown in bellow Figure 2.

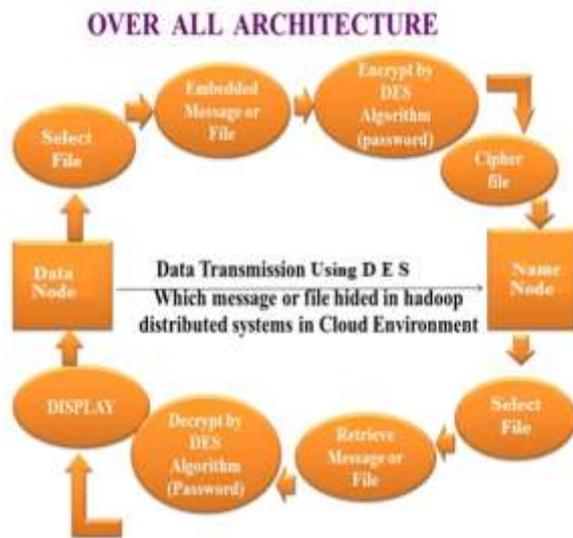


Fig. 2: Encrypted Data Transfer

IV. Methodology

This paper addresses the system of effective data transfer to the authenticated users and it incorporates a mechanism where in the unauthorized persons can be blocked from receiving the data, and among the authorized persons a security tag is attached so as to identify the source of data leakages. The users for whom the data is transmitted assuming to be authorized, if a data is leaked the security tag is set to 1 else the security tag is set 0. For all these security tags where the flag is one, the corresponding IP of the users will be scrutinized and an error warning will be notified. If the process is repeated the user or a person corresponding to particular data node with the particular IP will be blocked from receiving further data. In order to restrict unauthorized users to view the content the data is encrypted using DES and File key symmetric encryption algorithms. Symmetric encryption which is used in this paper is more advantageous than the Asymmetric encryption which requires more CPU cycles and CPU memory in addition to several limitations explained by kit File white et al [12]

V. Experimental Results

5.1 Environment Setup

Our experiments were performed on a 4 node cluster equipped with Hadoop. This project has been provisioned with one NameNode and four DataNodes. The NameNode was configured to use two 2.5-GHz CPUs, 2 GB of RAM, and 500 GB of storage space. Each DataNode was configured to use two 2.5-GHz CPUs, 2 GB of RAM, and 500 GB of disk storage. Besides this, all the computing nodes were connected by a gigabit switch. BOSS GNU Linux 4.1., Hadoop 0.20.1, and Java 1.6.0_6 were installed on both the NameNode and the DataNodes. For cloud set up we

used Xen hypervisor for virtualization and Eucalyptus for cloud infrastructure establishment.

5.2 Implementation

5.2.1 Mapper and Reducer

The mapper that contains my TMap algorithm is applied to every input data that has been transmitted in hadoop distributed environment. The data that is transmitted should be encrypted at each node with DES and Mapkey. The encrypted bundle of information will be stored at a common memory of hadoop called HDFS (Hadoop Distributed File System). The Client called datanode is allowed to read the information blocks if he is having authorization, otherwise tag value will be incremented by number of times he attempted to seize the information. When the user at a particular data node is having a tag value more than 0 will be recorded at log file of namenode so that the author is not allowed to access any further information in that cloud. That is the IP associated with that particular user will be blocked. This processing will be carried out in the Mapper and reducer whose job is to segregate all url's of a particular user so that temporal patterns of the user can be found. A typical algorithm with Map and reduce functions for identifying such temporal patterns is listed in the Table 1.

Table 1: TMap and TReduce Algorithm

Set the input path and the output path
<i>Step 1: Client selects the file at any datanode.</i>
<i>Step 2: Authentication using Mapkey then goto step 3.</i>
<i>Step 3: Check for the captcha, if captcha is not matched and tag value is greater than 0 then goto Step 10 else goto step 4. // Start of Map Function.</i>
<i>Step 4: Map (key, value)</i>
<i>Step 5: Client can send or receive data</i>
<i>Step 6: Client encrypts file by using Encryption algorithm (DES) and Map key by giving Authentication (Password).</i>
<i>Step 7: The cipher file is transmitted over the cloud through hadoop HDFS.</i>
<i>// Start of Reduce Function</i>
<i>Step 8: Reduce (key, value).</i>
<i>Step 9: If at any IP, any unauthorized user is attempted to access the data, tag value will be set and the log file of name node will be updated so that the IP will be blocked.</i>
<i>Step 10: If the IP is with authorized user then decrypt data by giving password.</i>
<i>Step 11: The message or file is retrieved at any data node of the respective cloud.</i>
<i>Step 12: Logout from datanode.</i>

Mapkey Algorithm:

The Mapkey algorithm is one of the security algorithms used to provide security for the user data and store them in an encrypted format. A random number is generated using Password Based Key Derivation Function (PBKDF2) Algorithm that derives the key by applying SHA1,SHA256,MD5 etc., algorithms for the random generated number and again different algorithms like AES,DES etc., are applied by choosing the random number as a secret key. When a user uploads the file in a cloud, the security algorithms are applied over the user file and encrypt the file and then the cloud provides an encrypted output file. These files are stored in a cloud storage database, which also provide high level security to the cloud computing environment. A secret key is provided to the authorized user to access his files in the cloud environment.

MapKey Algorithm:

1. Start
2. Read user file from at any datanode
3. Generate Random Number (n) // e.g.:12345
4. Perform PBKDF2 Algorithm to derive the Key (k)
5. Return MapKey
6. Stop

PBKDF2 Algorithm:

Input: Pwd Password

S salt Function

Ic Iteration Count

Kl Key length in bits $(2^{32} - 1) * Hl$

Parameters: Prf \rightarrow HMAC Function

Hl \rightarrow Hash Function Digest System

Output: Mk \rightarrow Master Key (Mk)

Algorithm: if $(Kl > (2^{32} - 1) * Hl)$

Return Error and Stop

Initialize L $\rightarrow [Kl / Hl]$

$Q = Kl - (L-1) * Hl;$

For(i= 1 to l)

$Xi=0;$

$V0=S // int(i);$

For(j= 1 to Ic)

$Vj= HMAC(Pwd,Vj-1);$

$Xi=Xi XOR Vj$

Return $Mk = Xi || X2 || \dots || Xi // <0 \dots Q-1>$

As shown in the above TMAP algorithm, the data encryption standards are one of the security algorithms used to provide security to the user data and store them in an encrypted format. When a user uploads the file in a cloud, the security algorithms are applied over the

user file. These files are stored in a HDFS which is a common hadoop Storage area for all data nodes in hadoop distributed environment. A secret key is provided to the authorized users to access this information in the cloud environment.

The following Table 2 shows the results for the log file, collected at an educational institution of size 10GB.

Table 2: shows running time for different log file size

File Type	File Format	Size (GB)	Number of Nodes	Time (seconds)
Text	Original file	10	4	600.8745
			8	500.3254
	Compression low	8	4	500.6789
			8	400.5973
			4	458.1020
			8	300.9848
High	6	4	500.6789	
		8	300.9848	

Sample Logfile

```
192.168.5.154,http://www.youtube.com/watch?v=t
Q2wJmFAvIE&feature=watch-now-
button&wide=1
192.168.5.65,http://apache.techartifact.com/mirror/
hadoop/core
192.168.5.178,http://www.gutenberg.org/ebooks/43
00
192.168.5.185,http://en.community.dell.com/suppor
t-
forums/laptop/f/3518/p/19364022/19818152.aspx#1
9818152
192.168.5.68,http://mail.google.com/mail/?shva=1
192.168.5.148,http://www.mediafire.com/error.php
?errno=378&quickkey=21abt2p569mkf6y
192.168.5.68,www.google.com
192.168.5.81,http://ubuntuforums.org/showthread.p
hp?t=932001
192.168.5.168,http://espressomind.wordpress.com/
2009/02/01/keeper-linksys-wrt54g-cannt-use-
wireless-interface/
192.168.15.81,http://mvgr-alumni.org/home/
alumni-news
192.168.5.148,http://www.youtube.com/movie?v=k
_PvLgr1c0&feature=mv_sr
192.168.5.114,http://www.facebook.com/
192.168.5.168,http://www.facebook.com/login.php
192.168.5.161,http://www.experts-exchange.com/
Software/CYGWIN/Q_23631555.html
192.168.5.115,http://apache.techartifact.com/mirro
r/hadoop/core
192.168.5.118,http://www.youtube.com/movie?v=k
_PvLgr1c0&feature=mv_sr
192.168.5.198,http://www.facebook.com/
192.168.5.206,http://www.facebook.com/login.php
192.168.5.161,http://www.experts-exchange.com/
Software/CYGWIN/Q_23631555.html
```

After preprocessing of log file by hadoop mapreduce, we can identify the temporal patterns of the users we

logged at regular or irregular interval of time is shown below.

Table 3: shows authorized vs unauthorized users

IP Address	URL's	Login time	Logout time	Tag	User type
192.168.5.154	http://www.youtube.com	9:55 AM	10:20 AM	5	unauthorized
192.168.5.65	http://www.google.com	11:30AM	03:00 PM	1	unauthorized
192.168.5.178	http://www.gutenberg.com	04:30PM	05:40 PM	4	unauthorized
192.168.5.185	http://www.pratibha.net	9:55 PM	10:20 PM	0	authorized
192.168.5.68	http://www.gitam.edu.in	11:20AM	12:10 PM	2	unauthorized
192.168.5.148	http://www.mvgrce.edu.in	10:05AM	11:00 AM	3	unauthorized
192.168.5.120	http://www.sharekhan.com	10:25AM	11:05 AM	0	authorized

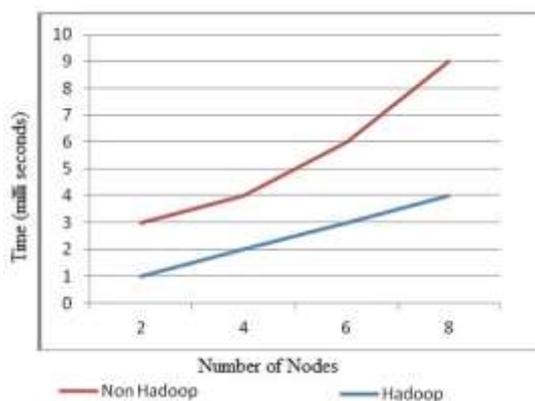


Fig. 3: shows running time vs number of nodes time complexity

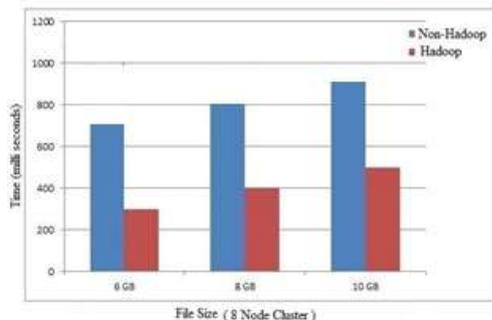


Fig. 4: shows usage of Hadoop Cluster in different file size (8 Node)

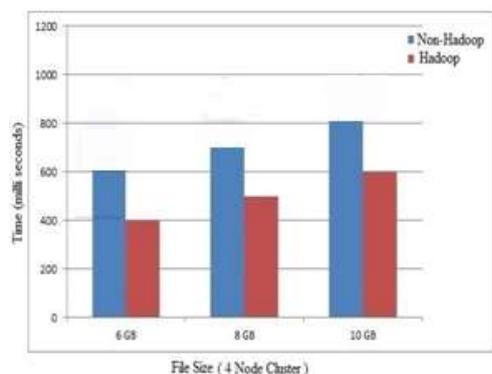


Fig. 5: shows usage of Hadoop Cluster in different file size (4 Node)

VI. Conclusion and Future Work

In this paper, an approach is presented to describe the security framework for cloud environment. This framework helps in providing the security to the user data in an encrypted format that are uploaded by the service user into a cloud, by incorporating the key features of different algorithms like DES, FileKey methods, which are placed in a Hadoop cluster. Key concepts of this architecture are the definition of unique security parameters for expressing security requirements and security functionality, we also provide a security system to the cloud environment by using the security functions and following the security parameters and security policies at the time of user login to provide authentication to the user and to find temporal patterns in the cloud. A novel approach of security mechanism can be extended in future work with the extension of more bit length of a key that are used in the different algorithms to provide more security to the user data in a cloud environment.

References

- [1] KejaingYe et al, vHadoop: A Scalable hadoop virtual cluster platform for MapReduce- Based Parallel Machine learning with Performance Consideration, 2012 IEEE International Conference on Cluster Computing Workshops, PP:152-160, 2012
- [2] J. Dean and S. Ghemawat, "Map Reduce: Simplified Data processing on large clusters", communications of the ACM. Vol.51, no.1,,pp 107-113, 2008.
- [3] T. White, Hadoop: The Definitive guide. yahoo press ,2010.
- [4] Mohamed H. Almeer et al "cloud hadoop mapreduce for remote sensing image analysis" Journal of Emerging Trends in Computing and Information Sciences, VOL. 3, NO. 4, April 2012.

- [5] Abhipal Singh et al “File Transfer Using Secure Sockets in Linux Environment”, Proceedings of the 4th National Conference; INDIACom-2010.
- [6] Matthieu Bloch et al “Network Security for Client-Server Architecture Using Wiretap Codes”, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, VOL. 3, NO. 3, SEPTEMBER 2008
- [7] Xinyi Huang et al “Further Observations on Smart-Card-Based Password-Authenticated Key Agreement in Distributed Systems”, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, 2013.
- [8] Bhaskar Prasad Rimal et al “A Taxonomy and survey of cloud computing system”, Fifth international joint Conference on INC,I MS, IDC, 2009.
- [9] Poulami Dutta et al “Data Hiding in Audio Signal: A Review” International Journal of Database Theory and Application Vol. 2, No. 2, June 2009
- [10] Elisa Bertino et al, “privacy-preserving digital identity management for cloud computing” IEEE computer society technical committee on data Engineering, 2009.
- [11] S.Fischer-Hubner and H. Hebdon, “PRIME privacy and identity management for Europe” August 2010.
- [12] kitu File whilte et al “symmetric vs Asymmetric encryption” 2010.
- [13] Kumar Gunjan et al,international journal of Engineering Research and Technology (IJERT),ISSN 2278-0181 vol1 issue4 june 2012.
- [14] Make D.Ryan et al “Cloud computing Security:The Scientific journal of systems and Software challenge, and a survey of solutions”,Elsevier,2013.
- [15] Jittin et al . ”An analysis on privacy preserving in cloud computing”, International journal of computer trends and Technology(IJCTT),volume 4 ,issue 6 june 2013.
- [16] Man Qi* et al, “Social Networking searching and privacy issues” information security technical report,Elsevier 2011.
- [17] Mohiuddin Ahmed et al, “An Advanced survey on Cloud Computing and state of the art Research issues” international journal of computer Sceince issues, vol 9 issue1,jan 2012.
- [18] P.Srinivasa Rao et al “A Novel and Efficient Method for Protecting Internet usage from Unauthorized Access Using Map Reduce” international journal of information technology and computer science, vol5, N3-6, Feb 2013.

Authors' Profiles



P.Srinivasa Rao currently working as Sr.Asst.Professor in CSE of MVGR College of Engineering. He is having Over 08 years of teaching experience. His research includes Data warehousing and Mining, Distributed Computing, Image Processing etc.



Dr. K. Thammi Reddy is the Director of Internal Quality Control(IQC) and Professor of CSE. at Gandhi Institute of Technology(GITAM).He is having Over 18 years of experience In teaching, Research, Curriculam Design and consultancy. His research areas include Data warehousing and Mining, Distributed computing, Network Security etc.



Dr. MHM. Krishna Prasad is the Associate Professor of CSE at JNTUK Kakinada, He is having Over 20 years of experience In teaching, Research, Curriculam Design and consultancy. His research areas include Data warehousing and Mining, Distributed computing, Computer Networks etc.

How to cite this paper: P.Srinivasa Rao, K.Thammi Reddy, MHM.Krishna Prasad,"A Novel Approach for Identification of Hadoop Cloud Temporal Patterns Using Map Reduce", International Journal of Information Technology and Computer Science(IJITCS), vol.6, no.4, pp.37-42, 2014. DOI: 10.5815/ijitcs.2014.04.04